Controlling the Output of Mechanized Retrieval Systems: An Empirical Test of the NASA *Thesaurus*

Submitted in Partial Fulfillment of
Requirements for Contract
NSR 39-011-076

From: University of Pittsburgh
Pittsburgh, Pennsylvania 15213

To: Chief, Dissemination Branch (Code UTD)
Technology Utilization Division
National Aeronautics and Space Administration
Washington, D.C. 20546

Submitted by: Allen Kent
Director
Knowledge Availability Systems Center

The substantive work involved in this report has been performed by Eleanor D. Dym, Donald L. Shirey, Martha Manheimer, and M. Ellen McAtee

Table of Contents

ABSTRACT

An experiment was designed and conducted to empirically evaluate the NASA Thesaurus. The evaluation focused primarily upon the Thesaurus functioning as a regulatory tool to control the output of a mechanized retrieval system. A ten percent random sample of search formula descriptors was used to perform a computerized search of the NASA file for the 4/67 search period. The resultant output became the standard against which the results of substituting thesaurally identified terms were compared. The figures of merit used in the evaluation were: redundancy, recall, and precision. The objective was to determine if the Thesaurus did function as a regulatory device even if the file had been indexed pre-thesaurally. Results were mainly negative. Effective control of output of searches against the NASA file was not attained by the thesaural modification of search formulas. The indexing of the pre-Thesaurus NASA file did not provide the degree of redundancy required to utilize efficiently the structural mobility afforded by the Thesaurus to regulate the output. This outcome suggests a follow-up study to determine whether post-thesaural indexing does indeed provide sufficient redundancy to permit formal use of the Thesaurus to formulate search prescriptions in which "broader" and "narrower" term relationships may be exploited effectively.

INTRODUCTION

The National Aeronautics and Space Administration (NASA) collects organizes, and indexes an extensive file of scientific-technical information related to the space effort. In order to facilitate the "spin-off" of information stored in the file, via innovations in products, processes, procedures, materials, and systems, to engineers, technicians, administrators etc., NASA is supporting six regional dissemination centers (RDC) throughout the United States. One RDC is administered by the Knowledge Availability Systems Center (KASC) at the University of Pittsburgh. Here a mechanized dissemination service is employed to exploit (i.e., identify, evaluate, and disseminate potentially useful information from) the NASA file on both a retrospective and a current awareness basis to industrial users.

Current procedures for exploiting the NASA file involve: (1) identifying and documenting the information needs peculiar to each user; (2) contructing a search formula based on the user's profile of information needs; (3) executing the computer search with this formula; (4) evaluating the search results; (5) providing the user with abstracts of those documents an analyst determines to be potentially relevant; (6) evaluation of the abstracts by the requestor in terms of relevancy to the question posed; and (7) controlling the system output by adjusting the search formula based on feedback information collected in (6) above.

The cumulative experiences of evaluators of information systems suggest four fundamental reasons for low recall and precision in mechanized searches of large files. They are:

A. The indexing was not consistent with the wide range of user interests. This is with respect to the identification of the different kinds of information the various users want to find in the documents.

B. The influence of the author's vocabulary on the nomenclature used by the indexers. Different terminology was used to describe identical or partially related concepts. Also the same terminology was frequently used to describe different concepts.

C. Differing depth (exhaustiveness) of indexing which led to uneven penetration into the potentially useful subject content of documents. Economic pressures and considerable variance in the indexer's understanding of the material in the document led to inconsistency in the number of independent concepts identified during analysis.

D. Deficiencies in the search formulas due to the uncertainties resulting from factors (A - C) above and uncertainties stemming from an incomplete understanding of what the system user really had in mind.

Recently a new tool, the NASA Thesaurus, has been made available to information specialists which should help to reduce these deficiencies in the indexing of documents and in the constructing of search formulas. It has several functions: first, it can serve as an authority list in the selection of legal terms for the construction of search formulas; second, it provides access to information files from any term entry point and by way of its cross-reference structure, channels the formula constructor to the correct entry term; third, it is used to regulate the system output with respect to both the quantity and quality of cited documents; and finally, it serves to remove ambiguity by showing the explicit interrelationships of the terms where such relationships exist.

Control of the output was generally achieved in two ways. One method was to change the logical operators connecting the terms of the search formula. This was considered a coarse adjustment (analogous to the coarse adjustment of a microscope). The fine adjustment then was the vocabulary that was employed in the formulas. The tool that was used to make the fine adjustment was a thesaurus.

The structure of the Thesaurus makes it useful in making the fine adjustments for controlling the output of a mechanized system. The three classes of structure generally found in thesauri provide the intrastructural mobility needed to reduce the previously discussed deficiencies in mechanized searches. The first class provides vertical mobility. Displays in this class are hierarchical, that is, a term and its related terms form arrays of conceptually broader and narrower terms that are the reciprocal one of the other. For example: the term "phase detectors" has as one related word the term "circuits" which is broader or more generic in meaning, and at the same time has the related word "synchroscopes" which is narrower or more specific in meaning. Substitution of the term "circuits" for the term "phased detectors" in a search formula should loosen the control of the output established by the original term, while substitution of the narrower term "synchroscopes" should tighten the control of the output by reducing it in size and scope. The second class of structure provides horizontal mobility. Provision of this type of mobility requires parallel displays where the terms are related conceptually or associatively, but not hierarchically. For example, "detectors," "phase control," "phase error," "synchronism," etc. are all related terms, all equally related, but in a parallel sense, not hierarchically. Included in this class are subject terms with broad or ambiguous meanings which precludes their being effective for indexing or retrieving information.

Such a term is "facilities." In this case the user was presented with a list of more specific terms to choose from, such as "airports," "buildings," "centers," "electrical power plants," etc. The third class of structure provided by the "use" and "used for" cross-referencing structure and was necessary to handle the synonymical relationships that might exist between terms. This mobility permitted entry into the system with an invalid term and assured access to the "legal" entry term.

Additional Background Information Necessary to Understand the Problem

Indexing of the NASA File.

Pre-thesaurus - Indexer generally identified the most specific set of terms which describes the document. He then did one or both of the following: a) disjoined the several words which comprised a pre-coordinated term, and/or b) posted up from the specific term to broader terms. Neither of these were done consistently. In addition, if other broader terms were considered legitimate index terms, they were also included.

The NASA system of indexing was never a "pure" uniterm system. There was always some provision made for the use of multiple-word, or precoordinate, terms that were used to index the documents in the published indexes. This was explained in a discussion of the NASA system by Van A. Wente.

> The NASA system designer recognized at the outset that these two approaches to retrieval, books indexes and tape files, may best employ two different types of indexing. It was felt that if the two types of indexing could be performed simultaneously and without duplication of effort, a significant gain in efficiency would be realized. On the one hand, the manipulative nature of computer tape data allowed the use of term coordination of an almost unlimited degree. On the other hand printed indexes, because of their non-

> manipulative nature, required the use of a *subject heading* type of indexing, whereby a highly developed structure of terms and cross-references guides the searcher to relatively more specific terms in the indexes where he might look for desired subject matter. In actual practice, the difference between these two indexing approaches became a difference in degree of what may be called *precoordination*.
>
> A completely word-by-word approach to indexing for computer retrieval may result in an excessive potential for false combinations and lack of specificity. Therefore, a machine vocabulary was developed which required precoordinations of those words most likely to give such results if used by themselves: proper names, specific things or projects and certain very general terms (LOW, HIGH, etc.). A vocabulary of approximately 13,000 terms resulted, with about 40 percent of its terms precoordinated. From 15 to 20 or more of these terms are currently employed to characterize the average document in the tape data. [1]

Each document acquired by NASA was, at the time of the sample period chosen for this study, assigned an average of three published terms and "10 machine terms," i.e., terms that were never used in the published indexes.[2]

The NASA system imposed a double burden on the analyst or user of the system in that he not only had to predict which term or terms might have been assigned to a document, he then had to be able to predict the form in which the term would appear, i.e., in an intact form as used in the published indexes or in the exploded form used on the machine-readable tape. If a multiple-word term had been used to index the document in the published indexes, it would appear on the machine-readable tape in the same form. In order to make the system consistent with uniterm principles, there was, at least part of the time, a policy that required that the multiple-word terms be fragmented into their component parts for use on the machine-readable tape.

[1] Van A. Wente, "Specificity and Accessibility in a System of Information Centers on Space and Aeronautic," *Colloquium on Technical Preconditions for Retrieval Center Operations*. ed. Benjamin F. Cheydleur (N. J. Macmillan, 1965) p. 56.

[2] This figure appears in the *SAL* in the Introduction and does not agree with the figure generated by the program used at the KAS Center.

Post-Thesaurus

In Indexing Guidelines for Use with the NASA Thesaurus, the following statement appears in the section labelled "Changes in Indexing Practices/Patterns":

> The unitizing (breaking down) of combined (precoordinated) published terms into machine retrieval terms will be discontinued.
>
> For example, LIQUID PROPELLANT ROCKET ENGINE (Old Subject Guide Term)
> LIQUID; PROPELLANT; ROCKET ENGINE or ROCKET and ENGINE
> (Machine Retrieval Terms)
>
> The Thesaurus will contain a modified version of most old "uniterms." These will be used based on the information being indexed and not automatically upposted as previously.[3]

Apparently "upposting" meant exploding the multiple word terms into their component parts.

This practice was evidently never firmly established as policy because it varied from time to time, and when it was in effect, was not consistently enforced. Therefore, Documentation, Inc., recommended that all strategies include both forms of the term, the exploded form because the document may not have appeared under the intact form in the published indexes because it was not a "major concept" in the document, and the intact form because if the document had been indexed under the intact form in the published indexes, the term may not have been exploded for use on the machine-readable tape.[4]

[3] National Aeronautics and Space Administration, "Indexing Guidelines for Use with the NASA Thesaurus, "Unpublished paper, January, 1968.

[4] Documentation Incorporated. Guide to the Processing, Storage, and Retrieval of Bibliographic Information at the NASA Scientific and Technical Information Facility, (College Park, Md., 1966) pp. 8-9.

Construction of Search Formulas and Searching Procedures

After the requestor's information needs had been determined, the set of keywords reflecting this need (his profile) were logically strung together with Boolean connectors to form a search formula. Search formulas can be sorted into two categories:

1. Analytics formed by a single term (representing a single aspect of information) or by a union of terms.

   The least complicated formula consisted of either a single term or a series of single terms summed. The union was represented in the examples by a plus sign (+). When this type of formula was used, it may be assumed that, in the opinion of the analyst, each term represents the question statement, and that every document indexed by the term had the possibility of being relevant to the question statement. The following are examples of strategies in this category.

   VACUUM DEPOSITION

   TEMPERATURE MEASUREMENT + THERMISTOR + TEMPERATURE SENSOR

   In the delimited population of 256 strategies used for this study, there were thirteen (5 percent) consisting of a single term and sixty-eight (27 percent) consisting of a series of two or more terms summed.

2. Analytics formed by the intersection of two or more search formula terms.

   The criterion for membership in this category was that two or more terms in the search formula were conjoined thus requiring the same set of terms to appear concurrently in a document (as index terms) for the logic of the analytic to be satisfied. Apparently in these cases, no single term represented the question statement and the analyst was interested in documents indexed by term "A" only if they were also indexed by term "B". The simplest formula of this type consisted of the intersection of two terms, for example,

OXIDATION * METAL

where the asterisk (*) was the logical connector for intersection. This type of formula most closely approximated the classic uniterm formula in which two three or four very broad terms, i.e., terms with non-specific referents, were conjoined in an intersection to retrieve document citations of a highly specific nature. Taube and Wachtel used, as an example, the coordination of the terms PLASTIC, BONDING, METAL and WOOD to find references to "plastic bonding metal to wood."[5]

More complex forms in this category of analytics are the following configurations which represent a series of intersections summed.

HYDROCARBON*(CATALYTIC ACTIVITY+CATALYSIS+CATALYST)

The following example is a variation of the preceding form.

((RADAR+ANTENNA)*(ARRAY+PHASED ARRAY))

Other formula variations consisted of the union of one or more terms with one or more intersections in various combinations. Examples are:

ADHESIVE + (ADHESION * POLYMER)

(FUSELAGE * (HELICOPTER + VIBRATION)) + HELICOPTER * VIBRATION)

A final variation is included which is a direct result of the formula constructor's uncertainty as to whether a multiple-word descriptor was disjoined or remained intact when selected as an index term.

As a result, analysts tended to reinforce their formulas by including both forms of the term. Four of the formulas contained a multiple word term as a single term and had its component parts used in an intersection.

[5] Mortimer Taube and Associates. Studies in Coordinate Indexing. 1953 (n.p. Documentation, Inc., 1953), II, 39.

For example

ELECTRIC MOTOR + (ELECTRIC * MOTOR).

However, in developing the list of terms used in the sample formulas no attempt was made to equate the terms forming intersections with a precoordinated form including both terms. A casual examination of the formulas will show that although some of the formulas include an intersection made up of two parts of a multiple-word term included in the Subject Authority List (SAL) and sometimes in the formulas in an intact form, many of the terms forming intersections were not represented in the SAL in a multiple word form.

Some formulas seemed to include both possibilities, that is, 1) the use of two terms to form an intersection not provided for in the terminology, and, 2) to back up a multiple-word term also used in an intact form in the formula.

The formula was then matched against the index terms of the documents in the file. If the terms matched and the requirements of the logic were met, the document was cited as potentially relevant to the question asked. The output was controlled both as to size and usefulness not only by the logic involved, but also by the selection of terms in the formula.

The SAL, which simply lists the legal terms without any indication of interterm relationships, was employed to control the vocabulary and to construct the population of search formulas used in this experiment (with the advent of the NASA Thesaurus, all of the search formulas have subsequently been modified to comply with thesaural requirements). In addition, the availability of the Thesaurus in 1968 exhibited relationships which provided varying degrees of intrastructural mobility, seemed to open a way for additional regulation of the output with fine control provided by vocabulary adjustments. All this can now be done with some *a priori* feeling as to what should happen to the output if one or several candidate terms are substituted for an original term in the formula.

Design of an Experiment

A. Objective: Empirically evaluate the NASA Thesaurus as to its capability of meeting its indicated use (i.e., regulate the quantity and/or quality of the output of searches executed against the subfile using search formulas constructed within its constraints).

Definitions (from the NASA Thesaurus)

1. Broader Term (B.T.) "... indicated that the terms that follow the B.T. notation represented more inclusive concepts that covered, among others, the term used. For example:

   ALUMINUM ALLOYS
   BT ALLOYS

2. Narrower Term (N.T.) "... indicated that the subject terms following the N.T. notation represented more specific concepts than the term used; it was a reciprocal of the broader term (B.T.) reference. For example:

   ALLOYS
   NT ALUMINUM ALLOYS

3. Related Term (R.T.) "... indicated that the two indexable terms were closely related conceptually but were not structured within the broader or narrower "tree," or hierarchy. The reciprocal of the R.T. reference "a" was the R.T. reference "b" and vice versa.

   a) RADAR EQUIPMENT
   R.T. RADIO EQUIPMENT

   b) RADIO EQUIPMENT
   R.T. RADAR EQUIPMENT

4. Analytics - Component parts of search formulas. Analytics were generally strung together with Boolean logic to form the basic tool for searching large files.

Definitions of Measures Used in the Experiment

5. Precision - Measures the proportion of the documents retrieved by the new analytics that were relevant to the user's information needs. This was not true precision in that the complement of the set (i.e., total cited by the analytics minus the relevant documents) was not evaluated by the user. It did, however, give an indication of the effectiveness of thesaurally substituting terms in search analytics when the standard of evaluation was the relevant documents retrieved by the original analytics.

6. Recall - Measured the ability of an analytic to retrieve relevant documents. There were two standards in this experiment against which recall was measured. One was the professional information analyst's opinion of the potential relevance of the document; the other was the information user's judgment of the document's relevance to his needs. New analytics formed with substituted terms (thesaurally suggested) were evaluated by their ability to recall the relevant documents retrieved by the original analytics.

7. Redundancy - The indexing of a document with a given term as well as any of the set of broader, narrower, or related terms that were thesaurally associated with the given term. It measured the co-occurrence of an index term with a thesaurally associated index term in the same document.

C. Expectations of Results: These fell into three classes based on the structural classification of the Thesaurus.

1. Broader term expectation: If a term structured by the Thesaurus as a broader term (B.T.) was truly a broader term, then the results of substituting this class into the search formulas should have been an output that contained all the documents indexed by the original search formula term in addition to others. The number of documents cited should be greater, but it must contain at least those cited by the original term (O.T.). Otherwise, the substitution degrades the search formula. This expectation is illustrated in Figure 1.

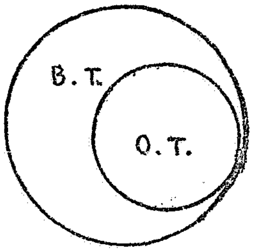


Figure 1

2. Narrower term expectation: If a term labeled a "narrower" term (N.T.) was truly a narrower term, then the results of substituting it into a search formula should have been an output that was a subset of the output from the O.T., but with increased precision. That is, a greater proportion of the cited documents should be relevant to the user's needs than would be in the original output. This expectation is illustrated in Figure 2. No document should be cited by the N.T.

that was not cited by the O.T.

In addition, the summation of all the documents cited by the narrower terms should be identical to the set of documents cited by the original term.

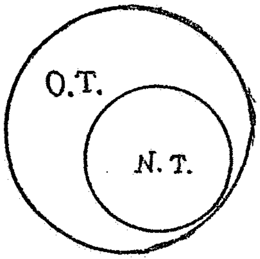


Figure 2

3. Related term expectation: If a term is a related term (R.T.) there are two potential results. First, the two listings could be disjointed (no documents in common), or second, if there is overlap, it should be slight. These two cases are illustrated in Figure 3.

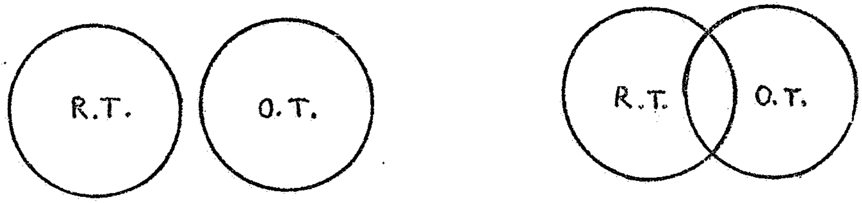


Figure 3

Redundancy with related terms should be low, but not zero.

Statement of the Problem

Can a thesaurus which has been based on an actual large file indexing vocabulary with its inter-term relationships established not by natural usage of these terms in the documents, but by conventions established by a panel of experts, be imposed over the construction of search formulas designed to search the same file as it was indexed pre-thesaurally as well as post-thesaurally? Can search formulas be thesaurally modified to regulate the output to the desired degree (with respect to recall and precision)? Do the terms in the thesaurus exhibit the relationships assigned to them (i.e., are "broader" terms really broader and "narrower" terms really narrower?

Methodology

The methodology described in this section is a summary of the important steps employed to design and conduct the experiment. Specific details on each of these steps are given in Appendix A.

1. Delimit a population of search formulas.

2. Make a listing of the unique terms in this population.

3. Equate the original terms (legal entrys in the pre-thesaurus Subject Authority List (SAL) with NASA thesaurus terms.*

4. Use the equated thesaurus terms as input to search the full thesaurus tape. Sort and list the mini-thesaurus of equated terms. The listing has the thesaurus term along with the subterms related to it.

5. Compare the Thesaurus terms with the SAL terms.

*When this study was first planned, it was believed that there would be a relatively small number of unique terms in the total number of terms used in the strategies and that it would be possible to visually equate the subterms listed under the strategy terms as Main Terms in the Thesaurus with the SAL form of the term and keypunch the SAL term for a single aspect search of the file. The unexpectedly large number of unique terms used in the strategies precluded this possibility and it was necessary to mechanize more of the project than had originally been planned.

6. Create a modified thesaurus containing only the original search formula terms and which has the NASA Thesaurus relationships and format, but uses SAL terminology.

7. Using the modified thesaurus, identify all the terms that are candidate terms for substitution with the randomly selected search formula term. These will include all the broader and narrower terms and approximately one-third of the related terms* for a given entry in the Thesaurus.

8. Keypunch the identified terms that are candidates for substitution as well as the original term into machine-readable form and search the NASA subfile for all documents indexed by these terms. This inverted subfile will be the data base for all subsequent searches.

9. Divide the candidate terms into two groups based on the type of analytic they form. The group containing the analytics formed by the union of one or more terms will be known as the $\bigcup$ group. The other group consisting of the analytics formed by the intersection of two or more terms will be known as the $\bigcap$ group.

10. For the $\bigcup$ group, thesaurally suggested terms will be considered a single aspect search formula and matched against the index terms of the subfile. The resulting list of accession numbers of documents indexed by this term will be compared to the listing of accession numbers for the original term and evaluated according to the criteria established for broader, narrower, and related terms (see expectations . . . page 13,14).

*Because of the large number of terms in the set of related terms, it was arbitrarily decided to use one-third of these as listed in the thesaurus.

11. For the $\cap$ group, an intermediate step must be taken before evaluating the structural relationship established for the selected term. The several constituent terms of the intersection must be searched simultaneously with the selected term. Thus, when a thesaural term is substituted, the set of terms comprising the intersection (including the thesaural term) must be searched as a unit. This unit is then considered a search formula and is matched against the file as was the single aspect formula. The results of these searches will then be evaluated as for the $\cup$ group.

Results and Discussion

A. Condition I: Substituting Thesaurally Broader Terms into Search Formula Analytics

1. Analytics formed by the original term either singly or in union with one or more terms in the search formula.

Central Tendencies and Dispersions

One hundred twelve (112) of the analytics that were randomly selected for study belonged to the $\cup$ group. Of these, broader terms were identified in the *Thesaurus* for sixty-five of the original terms in these analytics. The remaining forty-seven original terms had no broader terms identified for them. These original terms were index terms for a total of 390 documents in the subfile. The number of documents indexed by each term ranged from 0 to 43; the median indexed was three. For the 65 original terms there were 141 terms in the *Thesaurus* structurally classified as broader. These 141 broader terms in turn were index terms for 6506 documents.* The number of documents indexed by each broader term of this set ranged from 0 to 318; the median now being 4.9 documents per term.

* The number of unique documents may be somewhat less since no attempt was made to comb out the redundancy that occurs when a given document is indexed by two or more terms from the set of broader terms.

Redundancy

Our expectations for broader terms, if current search formulas can be thesaurally modified and used on the NASA file, is that there would be 100% redundancy, i.e., all 390 documents cited by the candidate terms would also be cited by the substituted terms. However, only 185 were cited giving a redundancy figure of 47.5%. Of the 185 redundant citations, 168 (or 91%) were the result of broadening the term by disjoining a multiple word descriptor (removing a modifier which acts to designate the specific subset of the several subsets of referents that may be attached to a given term, e.g., broadening a descriptor like "acoustic attenuation" by eliminating "acoustic." Thirty of the broader terms were also used as original terms in other component parts of the search formula, both in union with other terms and in intersection with other terms. Information analysts frequently employ this strategy not only to broaden the scope of the content area covered by the formula, but also to hedge against missing information because of their lack of faith in the consistency of the indexing and in the exhaustiveness of indexing. Sixteen of the thirty had been added in union (the Boolean "or") to the other term(s) in the formula. Of these sixteen, thirteen did not cite any of the documents cited by the candidate terms in the formulas. Again the expectation would have been that of 100% redundancy or overlap in citing the same document. The other three did provide some overlap, but the redundancy level was only 30%. This indicates that the level of redundancy added by thesaurally "broadening" the search formula is not satisfactory for the NASA file for this class of analytic.

Recall and Precision

Of the 390 documents cited by the original term about half (198 or 51%) were evaluated as potentially relevant by the information analyst. One hundred and four (104) of the 198 would have been cited had a broader term been used in the analytic instead of the original term. This is a recall factor of .53, indicating that about half of the documents evaluated as relevant by the analyst would have been retrieved had all the broader terms been included in the analytics of the 65 formulas in this group. Employing a single broader term in the formula analytic would have produced results ranging from 100% recall to zero recall. The former is likely to occur if the broader term used in the keyword is a multiple-word descriptor. The latter occurred frequently (18 of 141 broader terms (13%) cited no documents).

With regard to the user's evaluation of relevancy, there were 143 of the 390 originally cited documents evaluated as relevant (37%). Of the 143 relevant documents, 82 would have been recalled using broader terms in the component parts of the search formula instead of the original term. This is a recall factor of .57, indicating again that a little more than half the relevant documents would have been retrieved had broader terms been used in the analytics rather than the original term.

2. Analytics formed by the candidate term in intersection with search formula terms.

Central Tendencies and Dispersions

The remaining 117 analytics of the original 229 sampled for investigation were from the $\cap$ group. Of these, the Thesaurus identified thirty-seven as having a set of broader terms that could be substituted for the original terms in the analytics. The remaining eighty original terms had no broader terms thesaurally identified.

The 37 terms having a substitutable set of broader terms will be examined more closely now. It should be recalled from the discussion of procedures (see p. 18) that the $\cap$ group had to be handled differently than the $\cup$ group. An analytic in the $\cap$ group demands the cooccurrence of two or more terms to satisfy its logical requirements. In turn, this means that for documents to be cited it must have among its descriptors all the terms specified in the analytic. For example, the analytic

... electrodeposition*nonmetal* inorganic ...

requires a document to be indexed by all three of these terms in order for it to be cited (machine selected) as matching the information need of the request. This example is a stringent form of vocabulary control and is as likely as not to retrieve nothing. Most $\cap$ analytics require the cooccurrence of only two terms. These are frequently nested, i.e., one term is held constant, but the second term can be chosen from a set of alternatives. An example of this might be

... electron* (optics + microscope + diffraction) ...

Electron is a must index term, but the required second term can be any of the three inside the parenthesis. In effect there are three analytics displayed in this example --- electron and optics, electron and microscope, and electron and diffraction.

With this mind, we turn back to the discussion of the Ω group. There were 128 analytics formed with the 37 original terms. The median number of analytics thus formed was 2.1 for each term; the range was from 1 to a high of 13. The number of documents indexed by the set of terms specified in the 128 analytics was 91. The range for each set of terms was from 0 to 13, the median number indexed was 1.0 per document.

For these 37 original terms there were 70 terms in the Thesaurus structurally classified as broader. When these 70 broader terms were substituted (where applicable) into the analytics involving the original term, there were 489 documents cited as matching the requirements. The number of documents cited by the analytics after broader terms were substituted ranged from 0 to 156 per formula. The average number of cites per analytic was 3.8.

Redundancy

As stated previously, our expectations for broader terms is 100% redundancy. We expected all 91 of the documents cited by the original analytics to also be cited by the "broader analytics". However, only 23 were cited giving a redundancy figure of 25.3%. Of the 23 redundant citations, 16 (or 69.5%) were the result of disjoining a multiple-word descriptor as discussed earlier.

Recall and Precision

The recall and precision values that result when broader terms are used in a search formula can be determined. There are two standards upon which recall and precision can be based. First, there is the professional information analyst's evaluation of the documents, and second, the information requestor's evaluation.

Of the 91 documents cited by the original analytics 31 were considered relevant by the information analyst. Eight of these would have been cited had broader terms been used in the analytics instead of the term that was chosen originally. This is a recall factor of .26 and means that had all the broader terms been used in a component of the search formula instead of the original term, roughly one-fourth of those documents that the analyst felt to be relevant to the request would have been retrieved. The recall factor for relevant documents (those the requestor himself evaluated as relevant) was nearly the same. Twenty-seven of the 91 originally cited documents were evaluated as relevant by the user. Seven of the relevant documents were cited by analytics containing a broader term in structuring this component of the search formula.

The relationship between term broadness (as empirically defined) and the redundancy value (as empirically measured) is graphically displayed in Figure 4.

There were 110 terms considered truly "broader" by the empirical definition. Forty-three of these had a redundancy factor

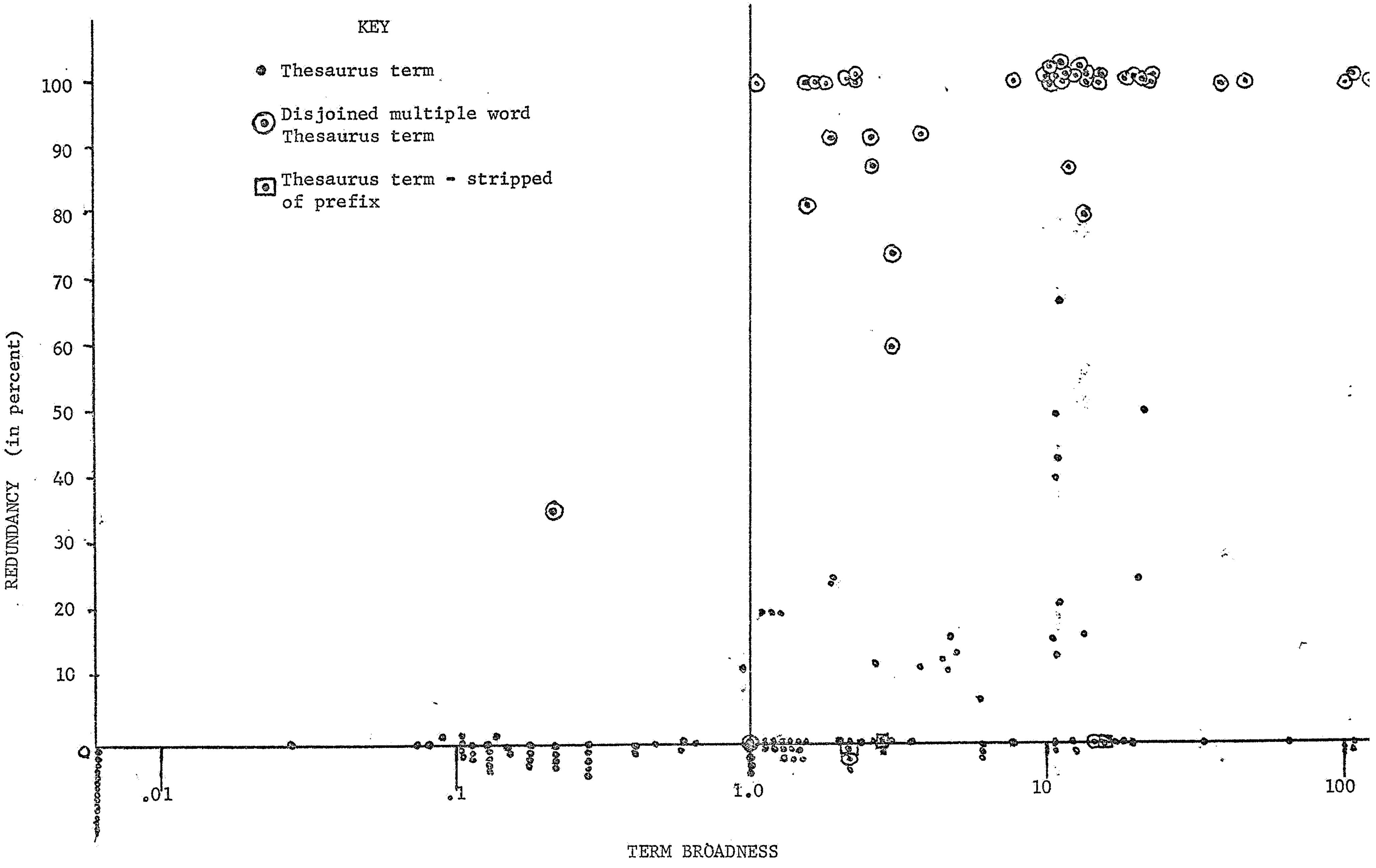

KEY
● Thesaurus term
⊙ Disjoined multiple word Thesaurus term
▣ Thesaurus term - stripped of prefix
REDUNDANCY (in percent)
100
90
80
70
60
50
40
30
20
10
0
.01
.1
1.0
10
100
TERM BROADNESS

Fig. 4

greater than 50%. In this set, all but four were candidate terms stripped of a modifier; 31 of the 43 had a redundancy factor of 100%. Of the remaining 67 (broader but with less than 50% redundancy) only six were disjoined multiple word terms or terms with the prefix stripped off (e.g. spectrophotography becoming photography). All six had redundancy factors of zero. Fifty-nine of the terms were not really "broader" by this definition (to the left of the line demarcating unity at 1.0). Of these 59, only five showed any degree of redundancy at all and only one of these is noteworthy (at 35%). It also happens to be a disjoined multiple-word descriptor.

Empirically, the substitution of broader terms into search formulas, the component parts of which were the union of terms, did increasethe number of documents cited by the substituted terms (approximately 16.5 times) when compared to the number cited by the candidate terms. However, the redundancy figure of 47.5 percent is not sufficient to fulfill the empirical definition of "broadness."

B. Condition II: Substituting Thesaurally Narrower Terms into Search Formula Analytics

I. Analytics formed by the original term either singly or in union with one or more terms in the search formula

Central Tendencies and Dispersions

One hundred twelve (112) of the analytics randomly selected for study belonged to the $\cup$ group. Of these, narrower terms were identified in the Thesaurus for forty-four of the original terms in the search formula analytics. The remaining sixty-eight of the original terms had no narrower terms identified for them. The original terms under consideration were index terms in a total of 529 documents in the subfile. The number of documents indexed by each original term ranged from 0 to 73, the median indexed was 7.5 documents.

For these 44 original terms, there were 200 terms in the Thesaurus classified structurally as narrower. These 200 narrower terms, in turn, were index terms for 632 documents.[*] The number indexed by each narrower term in this set ranged from 0 to 57, with the median being 2.8 documents.

Redundancy

Our expectation for narrower term substitution (as discussed under Design of the Experiment, pp. 13,14) is increased precision. The output for each narrower term substituted is a subset of the output for the original term. Therefore, the union of all the narrower term outputs

[*] Footnote on page 19 also applies to narrower terms.

should be the equivalent of the output for the original term. In addition, at least one of the narrower terms in a set should have high recall and precision values, indicating a convergence of user profile with the output. Therefore, we expected that all of the 529 documents cited by the original terms would be found in the 632 documents indexed by the narrower terms. However only 89 (or 17%) of the 529 documents were cited. In this case the sum of the parts equals only 17% of the whole, and not nearly meeting the expectation.

Recall and Precision

Of the 529 documents cited by the original term, about one-third (161 or 30.5%) were evaluated as potentially useful by the information analyst and forwarded to the user. Only 20 of the 161 analyst-relevant documents were cited by narrower terms, a recall factor of only .13. It indicates that a little more than 10% of the documents evaluated as potentially useful by the information analyst would have been retrieved had all the narrower terms been included in the analytics of the 44 formulas in this group. The use of a single narrower term in a formula analytic would have produced results ranging from 100% recall (one case and with only one document involved) to zero recall. The latter occurred over half the time (55%). Where there was recall with narrower terms, the recall value was .40 or less in each instance.

Considering now the user's evaluation of relevancy, there were 119 relevant documents out of the 529 mechanically cited with the original term; a relevancy factor of

21%. Of the 119 relevant documents, 16 would have been cited using all the narrower terms in the component parts of the search formula instead of the original term. This is a recall factor of 15%, a very low level even though the expectation was that all the narrower terms taken collectively would approximate the original term output.

2. Analytics formed by the candidate terms in intersection with search formula terms.

Central Tendencies and Dispersions

There were 117 analytics that belonged to the $\cap$ group. Of these, the Thesaurus identified sixty-eight as having a set of narrow terms that could be substituted for the original terms in the analytics. The remaining 49 original terms had no thesaurally identified narrower terms for them. There were 189 analytics formed by the 68 original terms in this group. The median number of analytics thus formed was 1.6 for each term and ranged from 1 to a high of 14. The number of documents indexed by this set of analytics was 376. The number for each term ranged from 0 to 56 with a median value of 1.9 documents per analytic.

For these 68 original terms there were 492 terms in the Thesaurus structurally classified as narrower. When these 492 narrower terms were substituted (where appropriate) into the analytics involving the original term, there were 192 documents cited as potential answers to the submitted questions. The number of documents cited by the new analytics after narrower terms were substituted ranged from 0 to 34 per formula. The average number of cites per analytic was 1.0.

Redundancy

As discussed previously, our expectations for narrower terms is that the collective output for the set of "narrower analytics" is equivalent to the output of the original analytics. We expected all 376 documents cited by the original analytics to be also cited by the summation of the narrower analytics. However, only 62 of the citations were redundant, giving a redundancy figure of 16.5%.

Recall and Precision

The recall and precision values, with respect to both the information analyst's and the user's evaluation of relevance that result from replacing terms in the search formula with narrower terms, also can be determined.

Of the 376 documents cited by the original analytics, 121 were considered relevant by the analyst. Fifteen of these would have been cited had narrower terms been used in the analytics instead of those originally chosen. This is a recall value of 12.5% and indicates that had all the narrower terms been used in the search formulas instead of the original terms only, one-eighth of the documents the analyst felt to be potentially useful to the request would have been retrieved. The recall factor for documents evaluated as relevant by the user was slightly less than for the analyst. Seventy-nine of the original cited documents were considered relevant by the user. Nine of these relevant documents were cited by analytics containing narrower terms substituted for the original. The recall for user-relevant documents is now 11.3% so that we now find that only one-ninth of the relevant documents can be recalled if the user happens to choose narrower terms in structuring this component of the search formula.

C. Condition III: Substituting Thesaurally Related Terms into Search Formula Analytics

1. Analytics formed by the original term either singly or in union with one or more terms in the search formula.

Central Tendencies and Dispersions

Analytics randomly selected for study that belong to the $\cup$ group numbered 112. Of these, related terms were listed in the Theasurus for 94 of the original terms in the search formula analytics. The remaining 18 analytics had no related terms associated with them.. The 94 original terms under consideration in this group were index terms for a total of 713 documents in the subfile. The number of documents indexed by each original term ranged from 0 to 73, the median number indexed was 4.2 documents.

For these original 94 terms there were 320 terms in the Thesaurus classified in its hierarchical structure as related. These 320 terms in turn were index terms for 5747 documents.* The number indexed by each related term in this set ranged from 0 to 367, the median being 2.8 documents.

Redundancy

Our expectation for related term substitution (discussed under Design of the Experiment, p. 14) is low redundancy, very low recall and precision. In many cases the set of documents cited by the original analytics and the set cited by "related analytics" should be disjoint; where there is overlap it should be slight.

* Footnote on page 19 also applies to related terms.

Therefore, we expected few (about 10%) of the 713 documents cited by the original terms to be among the 5747 documents cited by the related terms. This was correct, since only 128 (or 18%) of the 713 documents were cited in common. This amount of overlap did not differ greatly from the expectation. Thirteen of the 94 sets of cited documents were disjoint, citing no documents in common.

Recall and Precision

Of the 712 documents cited by the original term, one-third (238 or 33.4%) were evaluated as potentially useful by the information analyst and forwarded to the requestor. Related terms were found on 75 of the 238 analyst-relevant documents which is 31.5%. Roughly one-third of the documents evaluated as potentially useful by the information analyst would have been retrieved had all the sampled related terms been used in lieu of the chosen term for the 94 analytics in this group. When a single related term was used in a formula analytic, it resulted in recall values ranging from 100% (9 out of 9 documents) to zero. The latter occurred 30% of the time. Where recall was observed using related terms, it was found that the related term was a part of a disjoined, multiple-word descriptor, e.g., "gas" taken from "gas dynamics".

There were 178 documents in this group evaluated as relevant by the requestor. Of these, 51 could have been recalled using related terms from the sample in the analytics. The recall factor is .29.

2. Analytics formed by the candidate terms in intersection with search formula terms.

Central Tendencies and Dispersions

The number of analytics that belonged to the $\cap$ group was 117. Of these, 112 had a set of terms thesaurally identified for them which could be substituted for the original terms in the analytics. The other five original terms had no related term associated with them. There were 327 analytics formed by intersecting the 112 original terms with other terms. The median number of analytics thus formed was 2.0 for each term and ranged from 1 to a high of 14. The number of documents indexed by this set of original analytics was 550. The number for each analytic ranged from 0 to 95 with a median value of 1.0 documents per analytic.

For the 112 original terms there were 556 terms in the Thesaurus structurally classified as related. When these 556 related terms were substituted (again where appropriate) into the analytics involving original terms, there were 565 documents cited as potential answers to the submitted questions. After related terms were substituted to form new analytics, the number of documents cited ranged from 0 to 109 documents per formula. The average number of cites was 1.7 per analytic.

Redundancy

In an earlier discussion, we stated that our expectations for related terms was a low value of redundancy. We do not expect many of the documents cited by the "related analytics" to be in common with the set cited by the original analytics. This was indeed true, of the 550 documents cited by the original analytics, only 51 of them were also among the 565 documents cited by the related analytics. The redundancy figure of .09 was about what was expected.

Recall and Precision

Recall and precision values can also be determined for related terms when these are substituted into the search formula analytics. These values are based on the information analyst's and the user's evaluations of relevancy. Of the 550 documents cited by the original analytics, 144 were considered relevant by the analyst. Twelve of these would have been cited had related terms been used in the analytics instead of those originally selected. The recall value of 8.5% indicates that had all the related terms been used in the search formulas to replace the original terms, only one-twelfth of the documents the analyst felt to be potentially useful would have been retrieved. The recall factor for documents evaluated as relevant by the user was very low. Ninty-nine of the originally cited documents were considered relevant by the user. Four of these user-relevant documents were cited by analytics containing related terms in lieu of the original terms. Recall is now down to 4%, so that the prospect of retrieving relevant documents using related terms in search analytics is too dim to be useful.

Summary

The results of the experiment can be summarized in the following four tables. Table 1 is the redundancy factors for each of the three types of terms as they occurred in each of two component parts of search formulas. Table 2 is the recall values for substituting the three types when the information analyst's evaluation of relevancy is the criterion. Table 3 is the recall values when the information user is the judge of the relevancy. Table 4 is the precision values for the three types of terms when document relevancy (user) is the basis of comparison. Two values are given, one for the original term/analytic and one for the substituted term/analytic.

Table 1. Redundancy Factors for Thesaural Substitutions into Search Formula Analytics

	Relationship of Analytic in Search Formula	
Term type Substituted	Union	Intersection
Broader	.48	.25
Narrower	.17	.17
Related	.18	.09

Table 2. Recall Values for Thesaural Substitution into Search Formula Analytics: Analyst's Evaluation of Document as Criterion

	Relationship of Analytic in Search Formula	
Term type Substituted	Union	Intersection
Broader	.53	.26
Narrower	.13	.13
Related	.32	.09

Table 3. Recall Values for Thesaural Substitution into Search Formula Analytics: User's Judgment of Document as Criterion

	Relationship of Analytic in Search Formula	
Term Type Substituted	Union	Intersection
Broader	.57	.26
Narrower	.15	.11
Related	.29	.04

Table 4. Precision Values for Thesaural Substitution into Search Formula Analytics: Relevant Documents Cited by Original Analytics as Criterion

	Relationship of Analytic in Search Formula	
Term Type Substituted	Union	Intersection
Broader	.005 (.36)*	.015 (.30)
Narrower	.03 (.23)	.05 (.21)
Related	.01 (.25)	.007 (.18)

* Values in the parentheses are for the original analytics

Improvement in Search Formula Results

A considerable number of desired documents that were not retrieved by analytics containing the originally sampled terms would have been retrieved had a thesaurally related term been used in their places. The frequencies of occurrence of this type of event are given in the following table.

The Frequency of Improved Recall of Relevant Documents for Two Classes of Relevance Assessments

Type of Term	Analyst Relevant	User Relevant
Broader	34 (.09)	27 (.07)
Narrower	12 (.11)	11 (.10)
Related	18 (.06)	12 (.04)

Where values in the parentheses are precision values, i.e., proportion of total documents cited by the new analytics that wera relevant.

While thesaurally substituted analytics did retrieve a number of documents that would not have been retrieved by the original analytics, they also retrieved large numbers of non-relevant documents as indicated by the precision values. In most cases 90% or more of the retrieved documents using the analytics formed by thesaural substitutions were of no use. It can also be seen from the table that narrower terms did provide the best precision values, but only by very small margins.

CONCLUSIONS

The results of the experiment show that the thesaural modification of search formulas was, in general, ineffective in controlling the output of searches conducted against the large NASA file. This was so even though the Thesaurus was derived from the indexing vocabulary of the same file that was searched. The inter-term relationships found in the Thesaurus had been established by conventions that were agreed upon by a panel of experts rather than by the natural usage of these terms in the texts proper. But this is the lesser part of the answer, the greater part lies in the indexing.

The lack of effectiveness in regulating the output was observed to be true of the pre-thesaurally indexed NASA file, but may also apply to the file as indexed post-thesaurally. A visual examination of the post-Thesaurus indexing suggests to this writer that indexing with the aid of the Thesaurus is not significantly different from the pre-thesaural indexing under the NASA Subject Authority List (SAL). This, however, needs to be demonstrated empirically in a companion study to this experiment. If it is indeed true that the post-thesaural indexing is not significantly different from the SAL indexing which was shown to be lacking in the necessary redundancy for fine control of the retrieval procedures, then the only effective function being performed by the Thesaurus is the minor function of an authority list for the selection of descriptors. The major function as a tool for regulating the output of large mechanized retrieval systems remains unusable. And as a consequence, recall and precision values are likely to remain low.

We were unable to modify the bulk of the search formulas thesaurally and thereby regulate the output to a degree which in any way would be useful. Most of the difficulty can be attributed to the lack of a workable interface between the Thesaurus and the indexing which would permit the structural mobility incorporated into the Thesaurus to be imposed over the unstructured indexing found in the document file. One possibility for achieving this interfacing is to assign a value to each term in the Thesaurus based on accumulated search data from this and similar research experiences. The numeric value associated with each Thesaurus term would be an indicator of the effectiveness of each term in retrieving relevant documents rather than a frequency count of descriptor occurrences in the total file.

In an empirical sense, broader terms were broader. They did index many times more documents than the original terms, but the low redundancy figures precludes their effectiveness in controlling the output. The low recall and precision values substantiate this statement. High redundancy, where it did occur, was almost entirely the result of an indexing policy (since discontinued) of disjoining multiple-word descriptors.

Narrower terms were even less useful in controlling the output of searches. We expected the summation of the outputs of all the narrower terms for a given term to be equivalent to the output for that term. That is, we expected the two outputs to be 100% redundant. They were in effect only 17% redundant.

Related terms performed about as expected-low redundancy, low precision and recall.

The performance of thesaurally substituted terms in search analytics was disappointing with respect to the fine control of the output the *Thesaurus* is designed to provide. This is particularly devastating to those who construct search formulas for searching the NASA file not knowing the exact terminology the author is employing (which is also reflected in the descriptors selected by the professional indexer). The *Thesaurus* will not likely help him to home-in on the target documents with any fewer modifications than could be achieved by not using it. Information specialists will have to continue hedging their bets by stringing together all possible search analytics in logical union to guard against missing relevant documents.

Until the indexing of documents is performed in such a way as to conform to the structural relationships of the *Thesaurus*, an information specialist is just as well off using an authority list for selecting terms in a search formula as he is in using the *Thesaurus*.

Bibliography

Documentation Inc. Guide to the Processing, Storage, and Retrieval of Bibliographic Information at the NASA Scientific and Technical Information Facility. College Park, Md., 1966.

National Aeronautics and Space Administration, "Indexing Guidelines for use with the NASA Thesaurus", unpublished paper, January, 1968.

National Aeronautics and Space Administration, NASA Thesaurus: Subject Terms for Indexing Scientific and Technical Information. Preliminary Edition (NASA SP-7030), Office of Technology Utilization, Washington, D.C., 1967, 3 vols.

National Aeronautics and Space Administration, Subject Authority List, Unpublished list, April, 1967

Taube, Mortimer, et. al., Studies in Co-ordinate Indexing. Documentation, Inc., 1953.

Wente, Van A., "Specificity and Accessibility in a System of Information Centers on Space and Aeronautic," Colloquium on Technical Preconditions for Retrieval Center Operations. ed. Benjamin F. Cheydleur, New Jersey, Macmillan, 1965.

Appendix A

Specific Details of Methodology

1. Delimiting a population of search formulas.

For the single "current awareness" search period selected as the data base for this investigation, the indexes to 5600 "A" and "N" documents were added to the machine searchable file and more than eight hundred question statements were searched as part of the "current awareness" service of the Regional Dissemination Center at the University of Pittsburgh.

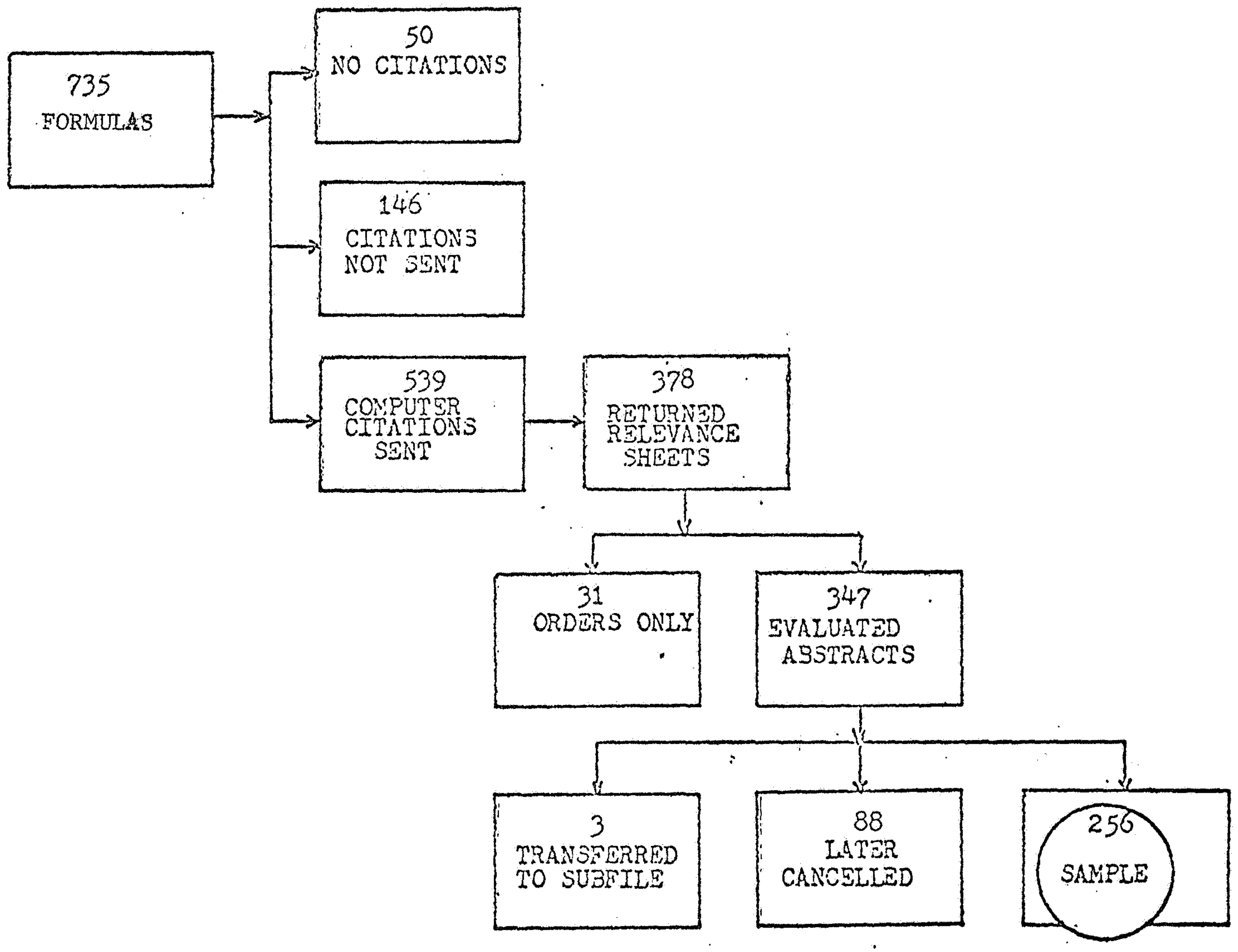
Since the usefulness of the Thesaurus would be determined in part by the user's evaluation of the cited documents, only those strategies for which this information was available were used. A second delimitation was that the question statement be represented by a unique search strategy. Almost two hundred of the question statements submitted to the system were not assigned unique strategies but were searched with similar question statements. These were not included in the sample. Approximately 7 per cent of the strategies searched retrieve document citations infrequently, primarily because their subject area is not one covered to any great extent in the aerospace literature. For this investigation, those strategies for which no documents were cited, as a result of the computer search and those for which none of the cited documents were forwarded to the user by the analyst were excluded from the sample. Because it was believed that it would be useful to compare the strategies developed after the publication of the Thesaurus with those in use before its effective date, strategies that were cancelled prior to January 31, 1968 were also eliminated from the sample.

Here, in summary, are the criteria used to select the sample of strategies forming the data base for this study, 1) the strategy must have been applicable to only one question statement; 2) the computer search must have resulted in citations some or all of which were judged relevant by the analyst and forwarded to the user; 3) the user must have returned a relevance sheet indicating the "relatedness" of the computer cited documents to his question statement, and, 4) the question must still have been in force after the effective date of the Thesaurus. The chart in Figure A-1 illustrated these delimitation.

2.* The SAL terms from the delimited opoulation of search formulas were keypunched, one term per card to form Deck A. These were then sorted alphabetically in order that the unique terms might be selected.

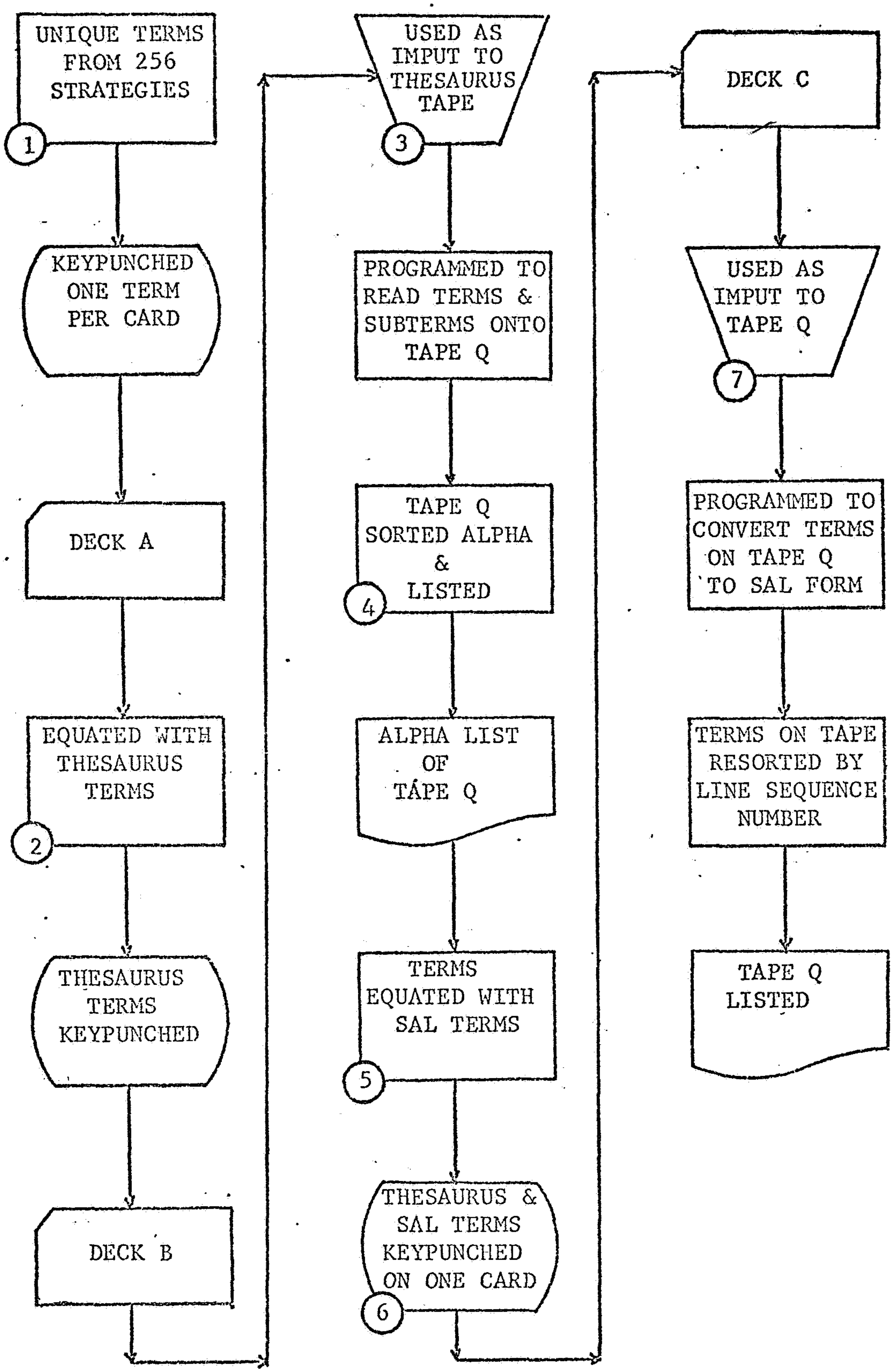
3. The unique terms were converted to the Thesaurus form of the term and the resulting list was keypunched, one term per card. This deck, Deck B, consisted of approximately 1500 terms. This set of terms does not exactly match the set of formula terms. In addition to those terms in the formulas for which relevance information had been returned by the user, the terms in all the formulas for which there was any feedback from the user relating to computer cited documents were included. Some of the returned relevance sheets included only document orders and were later deleted from the sample. Relevance sheets for several questions had not been returned by the users at the time the search was made of the files for the sample formulas. The terms for these formulas were later added to the sample but do not appear in the subsequent comparison of SAL and Thesaurus terminology.

*Circled numbers refer to appropriate parts of the schematic, figure A-2.



AN ANALYSIS OF THE FORMULAS SEARCHED DURING 4/67 SEARCH PERIOD

Figure A-1


UNIQUE TERMS FROM 256 STRATEGIES
1
KEYPUNCHED ONE TERM PER CARD
DECK A
EQUATED WITH THESAURUS TERMS
2
THESAURUS TERMS KEYPUNCHED
DECK B
USED AS IMPUT TO THESAURUS TAPE
3
PROGRAMMED TO READ TERMS & SUBTERMS ONTO TAPE Q
TAPE Q SORTED ALPHA & LISTED
4
ALPHA LIST OF TAPE Q
TERMS EQUATED WITH SAL TERMS
5
THESAURUS & SAL TERMS KEYPUNCHED ON ONE CARD
6
DECK C
USED AS IMPUT TO TAPE Q
7
PROGRAMMED TO CONVERT TERMS ON TAPE Q TO SAL FORM
TERMS ON TAPE RESORTED BY LINE SEQUENCE NUMBER
TAPE Q LISTED

SCHEMATIC OF PROCESS

Figure A-2

The Knowledge Availability Systems Center has a copy of the Thesaurus on computer tape. Deck B was used to read all of the strategy terms as Main terms along with their appropriate subterms from the computer tape of the _Thesaurus_. The terms were then written onto another tape.

The _Thesaurus_ is recorded on the computer tape as it appears in the first two volumes of the published _Thesaurus_, i.e., alphabetically by Main Term followed by subterms arranged alphavetically within the categories, "Used For," "Broader," "Narrower," and "Related."

One logical record equalled one 109 character record positioned in ten fields. Only four of the fields were necessary to this project and they were located in the following positions:

Field	Position	Content
I	1-7	a line sequencing number
II	8	a term relationship code
III	25-66	a 42 character subterm field
IV	67-108	a 42 character Main Term field

The Main Terms and the appropriate subterms were read from the taped _Thesaurus_ onto another tape by matching the terms in Deck B, the _Thesaurus_ form of the original strategy terms, with the terms in Field IV. When the term in Field IV matched the term on the card, the logical record was read onto another tape. When the terms no longer matched, the succeeding card was read from Deck B and the program continued to search in Field IV until all of the terms in Deck B had been matched with a term in Field IV and all of the subterms had been read onto another tape, Tape Q.

4. This second tape, Tape Q, comprising a subset of the Thesaurus, was sorted alphabetically by the terms in Field III, the subterms, and listed. This resulted in a list of approximately 22,000 terms of which 8,000 were unique.

5. Each unique term was then compared visually with the terms in the SAL

6. When a SAL term could be equated to the Thesaurus term on the alphabetized listing of Tape Q, both terms were keypunched on one card: the form found in the Thesaurus in the first 40 columns and the form found in the SAL in the second 40 columns. It was unnecessary to keypunch a card for those terms that exactly matched in the Thesaurus and SAL. When there was no term in the SAL that could be equated to the Thesaurus term, the Thesaurus term was keypunched in the first 40 columns and again in the second 40 columns of the card but prefixed in this case by a dash (-) in column 41.

This deck, Deck C, was used to substitute the SAL term for the Thesaurus term when there was a difference between the two forms, or to place a dash before the Thesaurus term if no equivalent term could be found in the SAL. After Tape Q had been modified by Deck C, Tape Q was re-sorted by line sequence number and listed. This provided the SAL form of the term in the display of terms afforded by the Thesaurus.

7. The deck, Deck C, that had been used to make the term substitutions on Tape Q was then used to create a deck (Deck D) for making single aspect searches of the sample period. The cards with a dash (-) in column 41 were sorted out of the deck since these terms did not appear in the SAL

The search program used at the University of Pittsburgh RDC requires that the search term begin in column one and limits the number of characters for any one term to twenty-one. A third requirement is that each strategy be preceded by a card which has a unique (to that run of the tape) number in the first five columns.

A program was written for the IBM Computer Model 1130 that read columns 41 through 61 on Deck C and punched a card with those characters in columns one through 21 on a new card. The IBM Model 1130 was also used to create the preliminary cards for the strategies, a deck of 2,500 sequential five digit numbers.

The IBM Computer Model 360-20 was used to interfile the term cards with the preliminary number cards, the search deck, and to list this deck, Deck D. This listing was necessary because the strategy terms do not appear on the computer printout of the search results. Only the unique number on the preliminary card appears and the only way to assign the printout of the search results to the appropriate term (strategy) is by the number used on the preliminary card.

There were two kinds of terms not in Deck C, the conversion deck, because it had been unnecessary to keypunch conversion cards for them.

1. terms that were exact matches in the SAL and Thesaurus did not need a conversion card because they remained the same on Tape Q.
2. access entries for terms appearing only as subterms on Tape Q did not appear on Tape Q and consequently no conversion cards were keypunched for them.

Terms that were exact matches in the SAL and Thesaurus were keypunched and added to the search deck.

The access entries for terms that were not Main Terms in the subset of the Thesaurus posed a greater problem. Each unique term appearing in the alphabetical listing of Tape Q had to be checked as a Main Term in the published Thesaurus in order to determine whether it had access (Used For) entries appearing as subterms. When this occurred, both terms were keypunched on the same card, the legal term in columns one through 38 and the access entry in columns 41 through 78. The unused columns were reserved for coding. This deck, Deck E, was sorted alphabetically by the access entry, beginning in column 41, and listed.

Alphabetizing by access entry was necessary to facilitate searching for the term in the SAL. When an equivalent term was found in the SAL, it was keypunched and added to the search deck.

Each search consisted of approximately one hundred and fifty single aspect strategies. After the search was completed, the strategy number cards and the term cards were separated on the sorter. The strategy number cards were then coded, beginning in column seven, and were reused in subsequent searches.

The deck of term cards was reproduced and the computer printout of the document accession numbers cited for the strategy was fastened to the appropriate card.

Deck C, the conversion deck on which were keypunched both the SAL and Thesaurus forms of the terms, and Deck E, the deck of access entries, were sorted manually and the terms categorized by the kind of difference between the SAL and Thesaurus forms of the terms.

All of the single aspect searches that were made were not used in this study. Only a very small number were used. The results of the other single aspect searches were used in another study. However, the complete process has been described because those used in this study were not searched separately but were part of the total searched.